

**PCT**WORLD INTELLECTUAL PROPERTY ORGANIZATION  
International Bureau

## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

<b>(51) International Patent Classification <sup>6</sup>:</b> <b>C12Q 1/68, G06F 15/18</b>	<b>A1</b>	<b>(11) International Publication Number:</b> <b>WO 96/38589</b> <b>(43) International Publication Date:</b> 5 December 1996 (05.12.96)
<b>(21) International Application Number:</b> PCT/US96/08115 <b>(22) International Filing Date:</b> 31 May 1996 (31.05.96) <b>(30) Priority Data:</b> 08/459,899 2 June 1995 (02.06.95) US <b>(60) Parent Application or Grant</b> <b>(63) Related by Continuation</b> US 08/459,899 (CON) Filed on 2 June 1995 (02.06.95) <b>(71) Applicant (for all designated States except US):</b> SMITHKLINE BEECHAM CORPORATION [US/US]; Corporate Intellectual Property, UW2220, 709 Swedeland Road, P.O. Box 1539, King of Prussia, PA 19406-0939 (US). <b>(72) Inventors; and</b> <b>(75) Inventors/Applicants (for US only):</b> STODOLA, Robert, King [US/US]; 309 Haws Lane, Flourtown, PA 19031 (US). TOBIN, Frank, L. [US/US]; 2964 Dorman Avenue, Broomall, PA 19008 (US). WILLIAMS, Arthur, L., Jr. [US/US]; 2924 Sunset Drive, Bethlehem, PA 18017 (US).	<b>(74) Agents:</b> BAUMEISTER, Kirk et al.; SmithKline Beecham Corporation, Corporate Intellectual Property, UW2220, 709 Swedeland Road, P.O. Box 1539, King of Prussia, PA 19406-0939 (US). <b>(81) Designated States:</b> JP, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).  <b>Published</b> <i>With international search report.</i>	
<b>(54) Title:</b> METHOD FOR ANALYZING PARTIAL GENE SEQUENCES		
<b>(57) Abstract</b>  A computer-based iterative method for analyzing partial gene sequences. Putative gene assemblies are built from partial gene sequences by the method. Incremental addition of new partial gene sequences to be integrated with an existing plurality of putative gene assemblies and a series of pre-processing steps prior to assembly allows efficient and accurate assembly of large amounts of partial gene sequences.		

**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	GB	United Kingdom	MW	Malawi
AT	Austria	GE	Georgia	MX	Mexico
AU	Australia	GN	Guinea	NE	Niger
BB	Barbados	GR	Greece	NL	Netherlands
BE	Belgium	HU	Hungary	NO	Norway
BF	Burkina Faso	IE	Ireland	NZ	New Zealand
BG	Bulgaria	IT	Italy	PL	Poland
BJ	Benin	JP	Japan	PT	Portugal
BR	Brazil	KE	Kenya	RO	Romania
BY	Belarus	KG	Kyrgyzstan	RU	Russian Federation
CA	Canada	KP	Democratic People's Republic of Korea	SD	Sudan
CF	Central African Republic	KR	Republic of Korea	SE	Sweden
CG	Congo	KZ	Kazakhstan	SG	Singapore
CH	Switzerland	LJ	Liechtenstein	SI	Slovenia
CI	Côte d'Ivoire	LK	Sri Lanka	SK	Slovakia
CM	Cameroon	LR	Liberia	SN	Senegal
CN	China	LT	Lithuania	SZ	Swaziland
CS	Czechoslovakia	LU	Luxembourg	TD	Chad
CZ	Czech Republic	LV	Latvia	TG	Togo
DE	Germany	MC	Monaco	TJ	Tajikistan
DK	Denmark	MD	Republic of Moldova	TT	Trinidad and Tobago
EE	Estonia	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	UG	Uganda
FI	Finland	MN	Mongolia	US	United States of America
FR	France	MR	Mauritania	UZ	Uzbekistan
GA	Gabon			VN	Viet Nam

METHOD FOR ANALYZING PARTIAL GENE SEQUENCESFIELD OF THE INVENTION

This invention relates to a computer-based method  
5 for building putative gene assemblies from partial gene  
sequences.

BACKGROUND OF THE INVENTION

The human genome is estimated to contain 3 billion  
10 base pairs of DNA. Within the genome, it is believed  
that approximately 50,000 to 100,000 gene coding  
sequences are dispersed. The gene sequences are thought  
to represent about 3% or approximately 90 million base  
pairs of the human genome.

15 It is generally recognized that elucidation of the  
structure of all human genes and their organization  
within the genome will be beneficial to the advancement  
of medicine and biology. Databases such as the Genome  
Sequence Data Bank and GenBank serve as repositories of  
20 the nucleotide sequence data generated by ongoing  
research efforts. Despite the efforts to date, GenBank  
lists the sequences of only a few thousand human genes.

Recent advances in automated, large-scale  
sequencing techniques have led to the initiation of two  
25 broad approaches to obtaining the sequence of the human  
genome. While the scientific debate continues as to the  
best approach, chromosome mapping and sequencing and  
gene sequencing projects have begun in earnest.

The Human Genome Initiative, a multinational effort having government backing in the United States and other countries, is attempting to characterize the genomes of humans and other model organisms on a chromosomal approach. In the private sector, large-scale sequencing of cDNA reverse transcribed from mRNA expressed in various human tissues, cell types and developmental stages is being pursued by a number of entities.

After publication of the Maxam-Gilbert and Sanger et al. nucleotide sequencing techniques, manual gene sequence assembly methods were practical for single gene or viral genome sequencing projects. As sequencing projects became more ambitious, manual techniques could be supplemented by computer-assisted sequence and contig assembly where overlaps between fragments were identified by software rather than by eye. However, the large scale of DNA sequencing projects and the rapidity with which sequence data is generated by automated sequencer machines has resulted in data analysis becoming a rate-limiting step in assembly of gene sequence data. The volume of data being generated by large-scale sequencing projects requires automated analysis in order to provide assembled sequence data in a timely manner.

Towards this end, efforts have been made to improve computer-assisted assembly of nucleotide sequence data. For example, in "Automated DNA Sequencing and Analysis", Adams et al. eds., Academic Press (1995), E.W. Myers

presents a discussion of software systems for fragment assembly in Chapter 32, while S. Honda et al. describe in Chapter 33 the Genome Reconstruction Manager, a long-term software engineering project to develop a system to  
5 support large-scale sequencing efforts.

Despite these efforts, a need exists for improvements over existing methods. The improved methods will provide computer-assisted nucleotide sequence assembly methods capable of more accurately and  
10 more efficiently assembling large amounts of sequence data.

#### SUMMARY OF THE INVENTION

Accordingly, one aspect of the present invention is  
15 a computer-based method for analyzing partial gene sequences. A computer-based iterative method for building putative gene assemblies from a plurality of partial gene sequences is provided. The method allows for the incremental addition of new partial gene  
20 sequences to be integrated with an existing plurality of putative gene assemblies. The method comprises preprocessing of the partial gene sequences and existing putative gene assemblies and assembling, responsive to grouping relationships, a consensus sequence from the  
25 preprocessed partial gene sequences and putative gene assemblies. Preprocessing comprises the steps of annotating regions within each of the plurality of partial gene sequences and each of the plurality of

existing putative gene assemblies; and grouping annotated partial gene sequences with other annotated partial gene sequences, where the other annotated partial gene sequences include components of the existing plurality of putative gene assemblies. Preprocessing allows for efficient and accurate assembly.

#### BRIEF DESCRIPTION OF THE DRAWING

10 The accompanying drawing, which is incorporated in and constitutes a part of the specification, illustrates a preferred embodiment of the invention and together with the description serves to explain the principles of the invention.

15 FIG. 1 is a block diagram of a method for analyzing partial gene sequences.

#### DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The method of the invention provides for automated management of a large and continuously growing population of partial gene sequences. As used herein, the term "partial gene sequences" refers to a series of symbolic codes for nucleotide bases comprising a portion of a gene, DNA or RNA. Partial gene sequences can be derived by automated or manual methods well known to those skilled in the art and can be stored in a database.

The method is a computer-based iterative process for building putative gene assemblies from a plurality of partial gene sequences. As used herein, "putative gene assembly" is an arrangement of partial gene sequences aligned relative to one another and combined to yield a consensus sequence. The iterative nature of the method of the invention allows for the incremental addition of new partial gene sequences to be integrated with an existing plurality of putative gene assemblies in an efficient manner.

Large numbers of partial gene sequences can be assembled by the method of the invention. The gene sequence assemblies produced by the method can be stored in a database and characterized for biological function. The nucleic acids represented by the gene sequence assemblies and the proteins the nucleic acids encoded are useful as drug discovery reagents and/or biomedical research tools.

As shown in FIG. 1, the method of the invention broadly comprises three steps of annotation, grouping and assembly. Efficient and accurate assembly of partial gene sequences is achieved through the assembly pre-processing steps of annotating and grouping and the use of the plurality of existing putative gene assemblies. The increased efficiency of the present method allows for high throughput of partial gene sequences.

Annotating is a process of identifying regions of partial gene sequences and putative gene assemblies that may cause two unlike sequences to be considered alike or otherwise produce inaccurate results in the grouping or assembly processes. These regions are likely to interfere with the correctness of the subsequent grouping and assembly steps of the method of the invention. The remaining unidentified regions are considered to contain useful information (for the purpose of grouping and assembly) and are used in the subsequent grouping and assembly steps. Regions identified as likely to interfere with subsequent steps are ignored in those steps.

Examples of regions which can be identified in the annotating step are sequences from species other than the one of interest and nucleic acids or DNA from cellular structures such as ribosomes and mitochondria. Low information regions which occur multiple times in a sequence such as polynucleotide runs, simple tandem repeats (STRs) and genomic repetitive sequences, such as ALU, can also be identified. Further, ambiguous regions and regions resulting from experimental error or artifacts are also identified.

After annotation, the annotated partial gene sequences are grouped with other annotated partial gene sequences. The step of grouping the annotated partial gene sequences is based on determining association relationships between an annotated partial gene sequence

and other existing annotated partial gene sequences, some of which may be components of previously identified putative gene assemblies. This process begins by ignoring the annotated regions from the partial gene sequences and previously identified putative gene assemblies. The partial gene sequences, with the annotated regions ignored, are then compared with the consensus sequence of previously identified putative gene assemblies, with the annotated regions ignored.

10 The partial gene sequences are also compared with each other, ignoring the annotated regions. The partial gene sequences are placed in groups based on the similarities found in these comparisons. Resulting groups thereby contain a collection of partial gene sequences that would appear to belong together, i.e., the grouping step produces a group of partial gene sequences that are thought to assemble together.

For each group from the previous step, the positional ordering of the partial gene sequences relative to one another is taken as a group on the assumption that all partial gene sequences belong to the same putative gene assembly. One of the consequences of the ordering may be that more than one putative gene assembly may result should the ordering step uncover inconsistencies among the group of partial gene sequences.

Once positional ordering has been completed for each putative gene assembly, a consensus sequence is

generated by a variety of contig assembly programs known to those of ordinary skill in the art. Exemplary is GELMERGE available from Genetics Computer Group, Inc. in Madison, WI.

5       The method of the invention is computer-based. Accordingly, partial gene sequences, annotated partial gene sequences, grouped annotated partial gene sequences and assembled consensus sequences are embodied as signals in a computer while being processed by the  
10   method of the invention.

      Upon completion of the annotating, grouping, and assembling steps, the putative gene assemblies are stored in a database. Putative gene assemblies may be characterized on the basis of their sequence, structure,  
15   biological function or other related characteristics. Once categorized, the database can be expanded with information linked to the putative gene assemblies regarding their potential biological function, structure or other characteristics.

20       For example, one method of characterizing putative gene assemblies is by homology to other known genes. Shared homology of a putative gene assembly with a known gene may indicate a similar biological role or function.

      Another exemplary method of characterizing putative  
25   gene assemblies is on the basis of known sequence motifs. Certain sequence patterns are known to code for regions of proteins having specific biological

CLAIMS

1. A computer-based iterative method for building putative gene assemblies from a plurality of partial gene sequences comprising the steps of:

5 (a) adding incrementally new partial gene sequences to be integrated with an existing plurality of putative gene assemblies;

(b) preprocessing the partial gene sequences and existing putative gene assemblies; and

10 (c) assembling a consensus sequence from the preprocessed partial gene sequences and putative gene assemblies.

2. The method of claim 1 wherein the preprocessing step comprises the steps of:

15 (1) annotating regions within each of the plurality of partial gene sequences and each of the plurality of existing putative gene assemblies; and

(2) grouping annotated partial gene sequences with other annotated partial gene sequences, wherein  
20 some of the other annotated partial gene sequences may be components of existing putative gene assemblies.

3. The method of claim 1 further comprising the step of:

(d) characterizing the consensus sequence.

25 4. The method of claim 3 wherein the characterization of the consensus sequence is on the basis of homology to known sequences.

characteristics such as signal sequences, transmembrane domains, SH2 domains, etc.

In addition to the methods just discussed, which can be automated, genes may also be characterized on the basis of expert commentary from relevant human specialists for given genes or by the results of biological experiments.

It will be apparent to those skilled in the art that various modifications can be made to the present method for analyzing partial gene sequences without departing from the scope or spirit of the invention, and it is intended that the present invention cover modifications and variations of the method for analyzing partial gene sequences provided they come within the scope of the appended claims and their equivalents.

5. The method of claim 3 wherein the characterization of the consensus sequence is on the basis of similarities to known sequence motifs.

6. A computer-based iterative method for building  
5 putative gene assemblies from a plurality of partial gene sequences comprising the steps of:

(a) adding incrementally new partial gene sequences to be integrated with an existing plurality of putative gene assemblies;

10 (b) annotating regions within each of the plurality of partial gene sequences and each of the plurality of existing putative gene assemblies;

(c) grouping annotated partial gene sequences with other annotated partial gene sequences, wherein  
15 some of the other annotated partial gene sequences may be components of existing putative gene assemblies; and

(d) assembling, responsive to grouping relationships, a consensus sequence from the grouped annotated partial gene sequences.

20 7. The method of claim 6 further comprising the step of:

(e) characterizing the consensus sequence.

8. The method of claim 7 wherein the characterization of the consensus sequence is on the basis of homology to known sequences.

9. The method of claim 7 wherein the  
5 characterization of the consensus sequence is on the basis of similarities to known sequence motifs.

10. A computer-based iterative method for building putative gene assemblies from a plurality of partial gene sequences comprising the steps of:

10 (a) adding incrementally new partial gene sequences to be integrated with an existing plurality of putative gene assemblies;

(b) annotating regions within each of the plurality of partial gene sequences and each of the  
15 plurality of existing putative gene assemblies;

(c) grouping annotated partial gene sequences with other annotated partial gene sequences, wherein some of the other annotated partial gene sequences may be components of existing putative gene assemblies;

20 (d) assembling, responsive to grouping relationships, a consensus sequence from the grouped annotated partial gene sequences; and

(e) characterizing the consensus sequence.

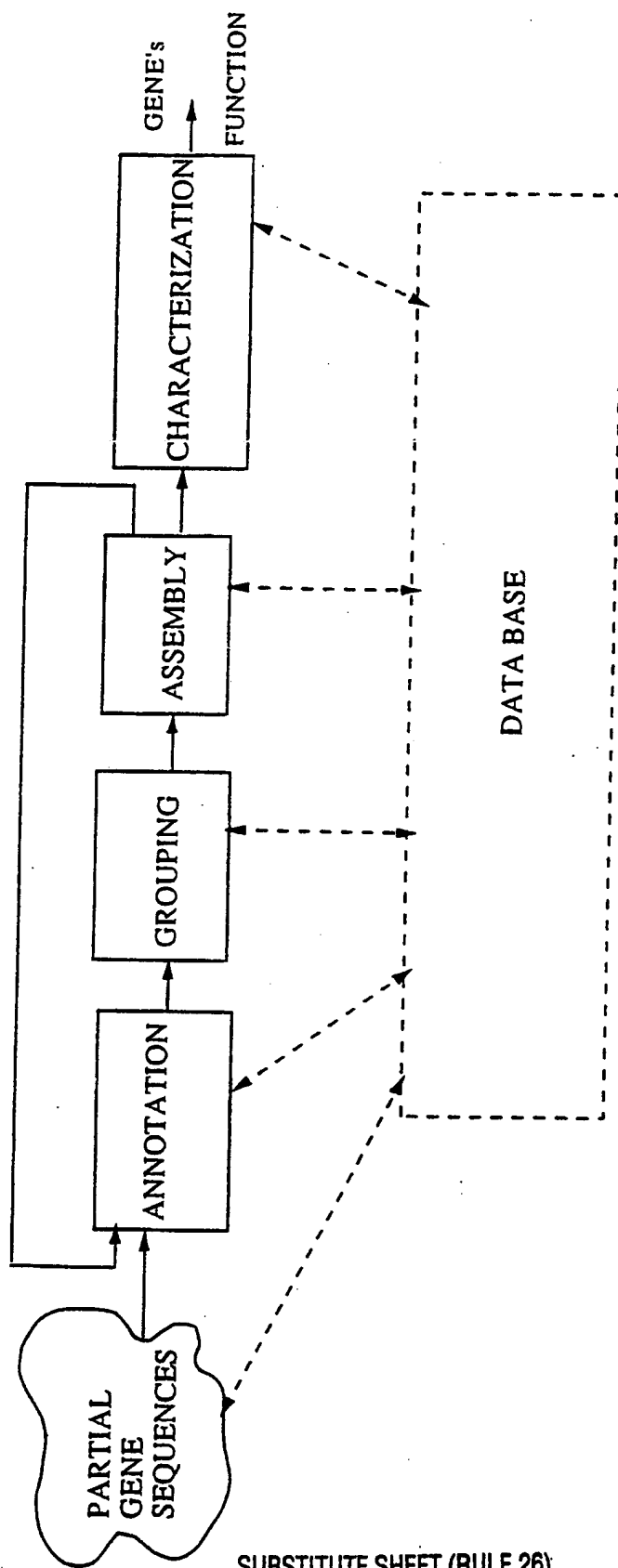


FIG. 1

# INTERNATIONAL SEARCH REPORT

International application No.  
PCT/US96/08115

## A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :C12Q 1/68; G06F 15/18

US CL :435/6; 395/13

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6; 395/13

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Please See Extra Sheet.

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	SMITH et al. The Statistical Distribution of Nucleic Acid Similarities. Nucleic Acids Research. 25 January 1985, Vol. 13, No. 2, pages 645-656, especially pages 646-650.	1-10
Y	PELTOLA et al. SEQAID: A DNA Sequence Assembling Program Based on a Mathematical Model. Nucleic Acids Research. 1984, Vol. 12, No. 1, pages 307-321, see entire document.	1-10
Y	GCG. Fragment Assembly. GCG Computer Group, Inc. 1994, pages 3.4-3.48, see entire document.	1-10

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:	*T	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
*A* document defining the general state of the art which is not considered to be of particular relevance	*X*	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
*E* earlier document published on or after the international filing date	*Y*	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
*L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	*Z*	document member of the same patent family
*O* document referring to an oral disclosure, use, exhibition or other means		
*P* document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search

24 JUNE 1996

Date of mailing of the international search report

09 JUL 1996

Name and mailing address of the ISA/US  
Commissioner of Patents and Trademarks  
Box PCT  
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

THANDA WAI

Telephone No. (703) 308-0196

## INTERNATIONAL SEARCH REPORT

International application No.  
PCT/US96/08115

## C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	KARLIN et al. 'Patterns in DNA and Amino Acid Sequences and Their Statistical Significance.' In: Mathematical Models for DNA Sequences. Boca Raton: CRC Press, 1989, pages 133-157, especially pages 134-136.	1-10
Y	WATERMAN, Michael S. 'Consensus Patterns in Sequences.' In: Mathematical Models for DNA Sequences. Boca Raton: CRC Press, 1989, pages 93-115, especially pages 94-95.	1-10
Y	WATERMAN, Michael S. 'Sequence Alignments.' In: Mathematical Models for DNA Sequences. Boca Raton: CRC Press, 1989, pages 53-92, especially pages 61-64.	1-10

# INTERNATIONAL SEARCH REPORT

International application No.  
PCT/US96/08115

## B. FIELDS SEARCHED

Electronic data bases consulted (Name of data base and where practicable terms used):

APS, JAPIO, EUROPEAN PATENTS, DERWENT WORLD PAT., DERWENT WPI, DERWENT  
BIOTECHNOLOGY

search term: genetic algorithm, computer, sequence, fragment, gene(s), automat?